NeoBiota 40: 73–86 (2018)
doi: 10.3897/neobiota.40.28840
http://neobiota.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

Undelivered risk: A counter-factual analysis of the biosecurity risk avoided by inspecting international mail articles

Sandy Clarke[1], Nyree Stenekes[2], Robert Kancans[2], Chris Woodland[3], Andrew Robinson[4]

1 *Statistical Consulting Centre, University of Melbourne, Australia* **2** *Australian Bureau of Agricultural and Resource Economics and Sciences, Department of Agriculture and Water Resources, Canberra, Australia* **3** *Compliance Division, Department of Agriculture and Water Resources, Canberra, Australia* **4** *Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Australia*

Corresponding author: *Sandy Clarke* (sjclarke@unimelb.edu.au)

Academic editor: *M. van Kleunen* | Received 3 August 2018 | Accepted 9 November 2018 | Published 4 December 2018

Citation: Clarke S, Stenekes N, Kancans R, Woodland C, Robinson A (2018) Undelivered risk: A counter-factual analysis of the biosecurity risk avoided by inspecting international mail articles. NeoBiota 40: 73–86. https://doi.org/10.3897/neobiota.40.28840

Abstract

International mail articles present an important potential vector for biosecurity and other regulatory risk. Border intervention is a key element in Australia's biosecurity strategy. Arriving international mail articles are inspected and those that are intercepted with biosecurity risk material are documented, including the address to which the article was to be delivered. Knowledge about patterns in the intended destinations of mail article permits more detailed biosecurity intervention. We used geo-location software to identify the delivery address of mail articles intercepted with biosecurity risk material from 2008–2011. We matched these addresses with demographic data that were recorded at a regional level from the Australian Bureau of Statistics 2011 Census and used random forest statistical analyses to correlate various demographic fields at the regional level with the counts of seized mail articles. The analysis of the seizure counts against demographic characteristics suggests a high correlation between having higher numbers of university students that speak a particular language in a region and higher quantities of intercepted mail articles destined for that region. We also explore metropolitan and regional patterns in the destinations of seized materials. These results can be used to provide information on policy and operational actions to try to reduce the rate at which mail articles that contain biosecurity risk material are sent to Australia.

Copyright *Sandy Clarke et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords
biosecurity, random forest, international mail, demographic analysis, population characteristics, Australian Bureau of Statistics 2011 Census of Population and Housing, spatial, prohibited goods, intercepted mail, risk profiling, compliance behaviour, data mining

Introduction

Biosecurity is the management of risks to the economy, the environment and the community, of pests and diseases entering, emerging, establishing or spreading (see, e.g. Craik W and Palmer D and Sheldrake R 2017). Australia's unique flora and fauna and strong reliance on agriculture make it specifically sensitive to invasive pressures, hence, many food, plant and animal products are considered to present a substantial biosecurity risk as they may be vectors for invasive pests. Its world-class biosecurity system is multilayered and complex, comprising activities undertaken offshore, at the border and onshore, by a broad range of participants that includes all Australian governments, industry, exporters and importers, farmers and other stakeholders (Craik W and Palmer D and Sheldrake R 2017). Australia's Department of Agriculture and Water Resources (hereafter, the department) plays an integral role in the biosecurity system. The department is the regulatory authority and inspectorate that is responsible for maintaining border biosecurity, amongst other things, focusing on biosecurity to protect agriculture, social amenity and the environment. It carries out this important responsibility via a suite of activities, including screening, monitoring, inspection and, when necessary, litigation. For example, the department sets out the conditions under which food, plant and animal products may be imported in the online Biosecurity Import CONditions database, BICON (Department of Agriculture 2018).

Incoming mail presents an important threat to biosecurity because it provides a pathway by which pests can enter (Meyerson and Reaser 2002). In 2014, approximately 176 million mail articles entered Australia, via one of four international gateway facilities, located in Sydney, Melbourne, Brisbane and Perth. The department performs both targeted and random inspections of incoming mail at each of these facilities, amongst other activities. Generally speaking, the intervention is carried out in order to intercept regulated pests, to verify the compliance of pathways with Australia's biosecurity legislation, to monitor the international environment and to detect and deter malfeasance. Screening and inspection at the gateway facilities may be by x-ray, specially trained detector dogs or opening the article and examining the contents. Mail items that are found to contain biosecurity risk material are seized and treated according to biosecurity policy (Department of Agriculture 2016a). The details of the seized item, including the location to which it was addressed, are recorded.

Here we report a counterfactual analysis of the factors common to the delivery addresses of seized mail articles, by exploring the relationship between population characteristics in these locations and interception data. The analysis is counterfactual in the sense that we are examining the patterns of biosecurity risk that was *not* delivered; the mail articles were seized upon arrival. The motivation for the analysis reported

in this paper was to better understand the patterns of destinations of seized materials and the characteristics of the regions of higher seizure risk. Prior to this analysis, there was some anecdotal evidence linking the importation of particular commodities, via the post particularly, with areas of large numbers of young people attending university. This project allowed a more rigorous assessment of this evidence, by investigation of the assumptions to see whether or not they are reflected in the historical data and to identify any other associations between population groups and the importation of high risk biosecurity goods, that are not yet known.

This knowledge would enable the department to target public relations campaigns that would focus upon informing the public about the risks and laws around the importation of foreign materials into Australia.

Methods

Data preparation

We used two sets of data to explore the relationship between the delivery addresses of seized articles and population characteristics or demographics of the local area.

First, the department's Mail and Passenger System (MAPS) database provided data about mail articles seized from 1 January 2008 until 31 December 2011, including a variety of characteristics of the article, the most relevant of which were the intended delivery address, which could be used for geocoding and the nature of the seized goods. The second set of data, provided by the Australian Bureau of Statistics (ABS), contained a range of demographic fields of consideration from the 2011 Australian Census. These were provided at the level of statistical unit 2 or SA2 polygons, one of the main structures within the Australian Statistical Geography Standard framework, a framework designed to enable consistent and comparable publication of statistics. SA2s are medium-sized geographical units intended to represent a community that interacts together socially and economically (Australian Bureau of Statistics 2016a). They are the smallest area for the release of many ABS statistics. There are 2,214 SA2 spatial units within Australia with populations ranging from 3,000 to 25,000. Visual comparisons as well as exploratory analysis at the finer grade of Statistical Area 1 (SA1) suggested that the SA2 was a more appropriate unit of analysis due to the increased number of seizures per unit, while not diluting the key patterns and relationships.

There is a large number of Census fields available for the analysis of individuals in the Australian population, therefore the analysis needed to focus on a smaller number of Census fields for the project to be feasible. *Profiles* were therefore developed of persons or groups associated with the admission of high risk biosecurity goods into Australia. Risk factors for the profiles were based on known intelligence and anecdotal evidence, provided by internal departmental studies, internal Customs studies and broader anecdotal evidence of links between imports and specific demographic groups

(Department of Agriculture 2016b, Maller C, Kancans R and Carr A 2007, Kancans R, Stenekes and Benedictos T 2010). In particular, this anecdotal evidence included a number of assumptions linking demographics and importation of particular commodities, in particular, that higher seizure incidences often occur in areas with large numbers of young people attending university.

Anecdotal and empirical research indicated that:

- culturally significant events may be the catalyst to the importation of high risk goods (plant and animal) through the postal system (e.g. cultural and religious festivals or events);
- culturally significant plants may be the catalyst to the importation of high risk plant material through the postal system;
- communication is a barrier to informing and educating persons of culturally and linguistically diverse (CALD) backgrounds about high risk biosecurity behaviour relating to the importation of plant and animal material through the postal system;
- cohorts of CALD persons may distribute biosecurity risk material due to close relational ties and spatial proximity;
- CALD persons studying in Australia are considered a high risk group; and
- CALD persons engaged in agricultural production could be a high biosecurity risk group due to the proximity to commercial agriculture.

The considerations discussed above led to a choice of ABS Census fields regarded as potential risk factors of individuals for the importation of biosecurity risk items which included:

- language spoken at home
- student attending an educational institution (university or other tertiary institutions)
- full-time or part-time attendance status of educational institution
- age in 5 year groupings
- people employed in industries related to horticultural production, including nursery and floriculture, mushroom and vegetable growing and fruit and nut tree growing.

In addition to these fields, the list of census fields was broadened to include those that had not been identified in the anecdotal evidence, but may still be important. These other fields are related to socio-economic circumstances including household characteristics, income levels and employment; those commonly used to describe a population. Land use within the SA2 was also a desirable field, due to the differing impact of biosecurity risk material for certain land uses, for example, primary production or parkland. This information was available by mesh block, the smallest geographic unit available from the ABS (Australian Bureau of Statistics 2016a). These units were aggregated to provide proportions of each land use type for each SA2.

The full set of census fields considered are included in Table 1. For the purpose of analysis, counts of each level of the census field were considered separately as

dummy variables, so there were technically 2854 variables used in the model to represent these 32 fields.

In order to be able to relate the seizure counts and these fields, we counted the number of seized articles that were addressed to locations within each SA2 polygon. This count required a match between the delivery addresses and the SA2 locations, which was performed using the G-NAF (Geographic National Address File) provided by Navigate[1]. This operation took consignee details addresses from items seized at the border and assigned geospatial parameters to it, namely the latitude and longitude and the ABS SA2 and SA1 areas. Overall, the geocoder worked as expected. Although failure rates were initially high, this was, however, more likely due to poor address data (e.g. special or foreign characters not readable by the G-NAF). This poor data quality was fixed through a thorough cleanse of the data. Once the geocoding was completed, this provided a key point of data that was unique across the MAPS data and the census information, namely the SA1 and SA2 locations. Using this standard data field, we were able to join and analyse the data with more accuracy.

Of the 404,873 seizure records in the period of consideration, 307,627 (76%) could be matched to a unique SA2 area. Failures to match were for a range of reasons, including that the international sender address was incorrectly included as the recipient address, that the address was poorly spelled or that the address was fake. Of the 307,627 matched records, 161,390 (52%) contained risk articles that either lacked appropriate Customs declaration or were misdeclared. These records were used for analysis – correctly declared items that were seized were not included.

These seizure counts per SA2 polygon were then matched with the equivalent demographic fields in order to explore the demographic characteristics of high seizure areas.

Random forest analysis

In seeking to determine which demographic fields are related to high seizure areas, one challenge is the very large number of possible demographic relationships. Therefore, a robust methodology was required to determine which were most closely related to increased seizures. A flexible model was also needed, as these relationships are not necessarily linear nor consistent across demographic fields. Tree-based models are valuable in exploratory cases such as these, because they allow for complex relationships between predictors and outcomes, including interactions between predictors in the way they affect the outcome (Hastie et al. 2009). We can create a tree that progressively splits each group of observations into two subgroups, based on the level of individual demographics fields. The choice of splits can be optimised automatically in software to best separate those with high and low risk of seizures. Note that the seizure outcomes used in these models was the rate per 100,000 of population, to allow for differences in the population in each SA2. The choices of splits yield useful information about the importance of the demographic fields in determining those areas with high levels of seizures. Complex interaction patterns are also naturally incorporated into tree models.

Table 1. Census fields considered in the analysis.

Category	Code	Census field
Language	LANP	Language Spoken at Home
	ENGLP	Proficiency in Spoken English/Language
	ENGP	Proficiency in Spoken English
Diversity	CITP	Australian Citizenship
	YARP	Year of Arrival in Australia
	BPLP	Country of Birth of Person
	ANC1P	Ancestry 1st Response
Education	TYSTAP	Educational Institution: Attendee Status
	TYPP	Type of Educational Institution Attending
	HEAP	Level of Highest Educational Attainment
Household	HCFMD	Family Household Composition (Dwelling)
	CDCF	Count of Dependent Children in Family
	STRD	Dwelling Structure
	NEDD	Type of Internet Connection
	MDCP	Social Marital Status
	VEHRD	Number of Motor Vehicles (ranges)
	NPRD	Number of Persons Usually Resident in Dwelling
	TEND	Tenure Type
	MV1D	Household One Year Mobility Indicator
	RNTRD	Rent (weekly) Ranges
	MRERD	Mortgage Repayments (monthly) Ranges
Income	HIED	Equivalised Total Household Income (weekly)
	HIND	Total Household Income (weekly)
	INCP	Total Personal Income (weekly)
Employment	INDP	Industry of Employment
	GNGP	Public/Private Employer Indicator
	OCCP	Occupation
	POWP	Place of Work
	LFSP	Labour Force Status
Age	AGEP	Age in single years
Gender	SEXP	Sex
Landuse	N/A	Land use based on Mesh Block Category

It is possible to construct one such tree model for these data but over fitting is a genuine risk. As there is a relatively large number of potential demographic predictors available when fitting the model, it is possible to estimate the risk for the existing data very well, but the model may not be generalisable to other pathways. For example, it may highlight spurious relationships due to only a few unusual observations in this particular data set. One way to mitigate this problem is to create many trees based on random samples of the data and average the results of these, so the results are not sensitive to the specific data available. These kinds of models are called random forests, as they involve many trees generated from random samples of the data. These kinds of models have been shown to have very high performance (Fernández-Delgado et al. 2014) and the number of SA2s makes this approach feasible in this case. The

longitude and latitude of the centroid of each SA2 polygon was also included to allow for and assess spatial correlations (Mascaro et al. 2014). The particular random forests used for this analysis involved 5000 trees, with the minimum size at each terminal node set to 5.

All analysis used the open-source statistical environment R (R Core Team 2013), along with contributed packages randomForest (Liaw and Wiener 2002), sp (Pebesma and Bivand 2005), ggmap (Kahle and Wickham 2013), rgdal (Bivand et al. 2014), gstat (Pebesma 2004) and ggplot2 (Wickham 2009).

Results

Description of seizures

Total seizure numbers for the period are reported for each of 12 broad commodity categories and counts of these seizures are given in Tables 2 and 3, separated by state and region. Region in this case is defined using the ABS remoteness structures, which are based on a measure of relative access to services (Australian Bureau of Statistics 2016b). In order to identify patterns of dependence between the address locations and category (that is, whether certain states/regions are associated with more or less of certain seizure types) the cell deviations that contribute to a χ^2-test were considered (Greenwood and Nikulin 1996). A cell deviation less than -2 or greater than 2 was considered evidence for a lack of independence; here we observe some deviations of 10 times this amount which would be highly unexpected if the categories and locations were truly independent. Based on the magnitude of these cell deviations, given in brackets, there were proportionally fewer Animal Products destined for Victoria (-28.9) compared with New South Wales (NSW) (10.1) and Queensland (29.3) and greater amounts of Plant/Plant Products (15.0), also compared with these states (-10.0, -12.8). Overall, there were greater amounts of Human Therapeutics destined for Victoria (21.9), Vegetable/Vegetable Products destined for NSW (8.5) and Contaminated Goods/Footwear/Packaging destined for Western Australia (16.1). There are increased numbers of Plant/Plant Product seizures destined for regional areas (19.4) compared with metropolitan areas (-8.6) and increased numbers of Fruit/Fruit Product, Vegetable/Vegetable Products and Mushroom/Fungi seizures destined for metropolitan areas (4.4, 4.6, 5.2) compared to regional areas (-9.4, -10.4, -10.8).

Random forest results

As the random forest approach involves the averaging of many tree models, no single tree or diagram can be used to display the relationship between the demographic data and seizure rates. However, it is possible to consider the effect of each key demographic

Table 2. Counts of seizures from 2008–2011 by broad commodity category and region.

Broad commodity category	Remote	Regional	Metropolitan
Animal Products	484	6553	40148
Biologicals	17	236	958
Contaminated Goods/Footwear/Packaging	84	1356	5655
Fruit & Fruit Products	63	1165	8741
Grains, Legumes & Nuts	100	1472	10537
Herbs & Spices	91	1284	7623
Human Therapeutics	136	2048	12451
Live Animals	1	27	119
Mushroom /Fungi	13	442	4333
Plant/Plant Products	700	11884	52200
Soil/Mineral Samples & Fertiliser	10	253	867
Vegetable/Vegetable Products	46	549	4950
Total	1747	27336	149183
Overall rate per 100,000 population	19.6	32.4	88.7

Table 3. Counts of seizures from 2008–2011 by broad commodity category and state/territory.

Broad commodity category	NSW	Vic	Qld	SA	WA	Tas	NT	ACT
Animal Products	19093	9619	10208	2723	3175	810	266	1291
Biologicals	420	350	230	58	107	25	8	13
Contaminated Goods/Footwear/Packaging	2411	1950	1033	471	926	122	41	141
Fruit – Fruit Products	4045	2787	1416	643	636	192	46	204
Grains, Legumes – Nuts	4827	3227	1795	751	983	216	50	260
Herbs – Spices	3303	2426	1526	627	618	268	64	166
Human Therapeutics	4931	5390	1968	787	995	258	70	236
Live Animals	52	46	19	8	13	5	0	4
Mushroom/Fungi	2035	1286	561	341	282	162	14	107
Plant/Plant Products	22816	19717	9192	4672	5502	1440	289	1156
Soil/Mineral Samples Fertiliser	405	333	160	82	98	25	8	19
Vegetable/Vegetable Products	2474	1430	680	320	391	109	19	122
Total	66924	49017	28819	11501	13764	3637	876	3728

NSW: New South Wales, Vic: Victoria, Qld: Queensland, SA: South Australia, WA: Western Australia, Tas: Tasmania, NT: Northern Territory, ACT: Australian Capital Territory.

field, averaged over the other fields, using overall measures of importance. These are a summary of impact of the splits based on each field. These values are only relative and have been standardised such that the largest value is set to 100. This gives an indication of the relative importance of each demographic field in the prediction of seizures, as compared with the relative importance values for the other demographic fields reported for the same model. Table 4 gives the relative importance values for the top 10 fields in a model based on all the available data. In each case the relationship was positive, with increasing seizure rate as the demographic field increased. Note that some language and ancestry fields have been deliberately obscured: Country A refers to an East Asian country for which Language A is the official language. These have been

Table 4. Top 10 relative variable importance measures.

Census field	Level	Variable importance
Equivalised Total Household Income	Nil income	100
Total Household Income	Nil income	91.8
Year of Arrival in Australia	2010	48.6
Country of Birth of Person	Country A	31.4
Educational Institution	Full time student aged 25+	27.1
Type of Educational Institution	University or other tertiary	20.6
Australian Citizenship	Not Australian	16.9
Ancestry 1st Response	Country A	14.9
Language Spoken at Home	Language A	14.5
Total Personal Income	Nil income	13.7

obscured in order to avoid potential labelling of social groups in society on the basis of race or ethnicity, gender, age etc. The point of the study was not to profile groups pejoratively, but to explore a method that can be used to assist in designing campaigns to raise awareness.

Overall, no household income, some relationship to Country A, recent arrival and university study dominate this list of important fields. These fields are also highly related, as international students are often recent arrivals from Country A without any household income. However, it is worth emphasising that the model did not consider counts of individuals with all of these characteristics at the SA2 level, only the counts of each within each SA2.

Given this relationship with students, it is worth considering the destinations of seizures in relation to universities. Figure 1 represents the seizure rates for each SA2 for the two capital cities with the most seizures, Sydney and Melbourne. In many cases the red hotspots are close to the locations of key university campuses.

As the units of study (SA2s) were spatially related, it is appropriate to comment on any spatial patterns observed. While considered in the model to assess spatial correlation, the longitude and latitude of the SA2 polygons did not feature highly in variable importance, with relative importance values of 0.4 and 0.2, respectively. This indicates that those SA2s with similar location did not, in fact, have similar seizures frequencies, after adjustment for demographic fields. A variogram of the residuals of the random forest model has been provided in the suppl. material 1 and also shows no evidence of spatial correlation.

This could be due, in part, to the spatial patterns in the seizure counts being strongly related to that of the predictors in the model, leaving no detectable residual spatial correlation. A simple linear regression analysis between seizure rate and the latitude and longitude of the SA2 indicated an association between seizure rates and longitude (but not latitude), suggesting this might be happening in part at least.

While the overall seizure rate is of primary concern, different kinds of seizures pose different kinds of risks and require different mitigation policies in response. As indicated in Tables 2 and 3, there are 12 broad commodity categories into which

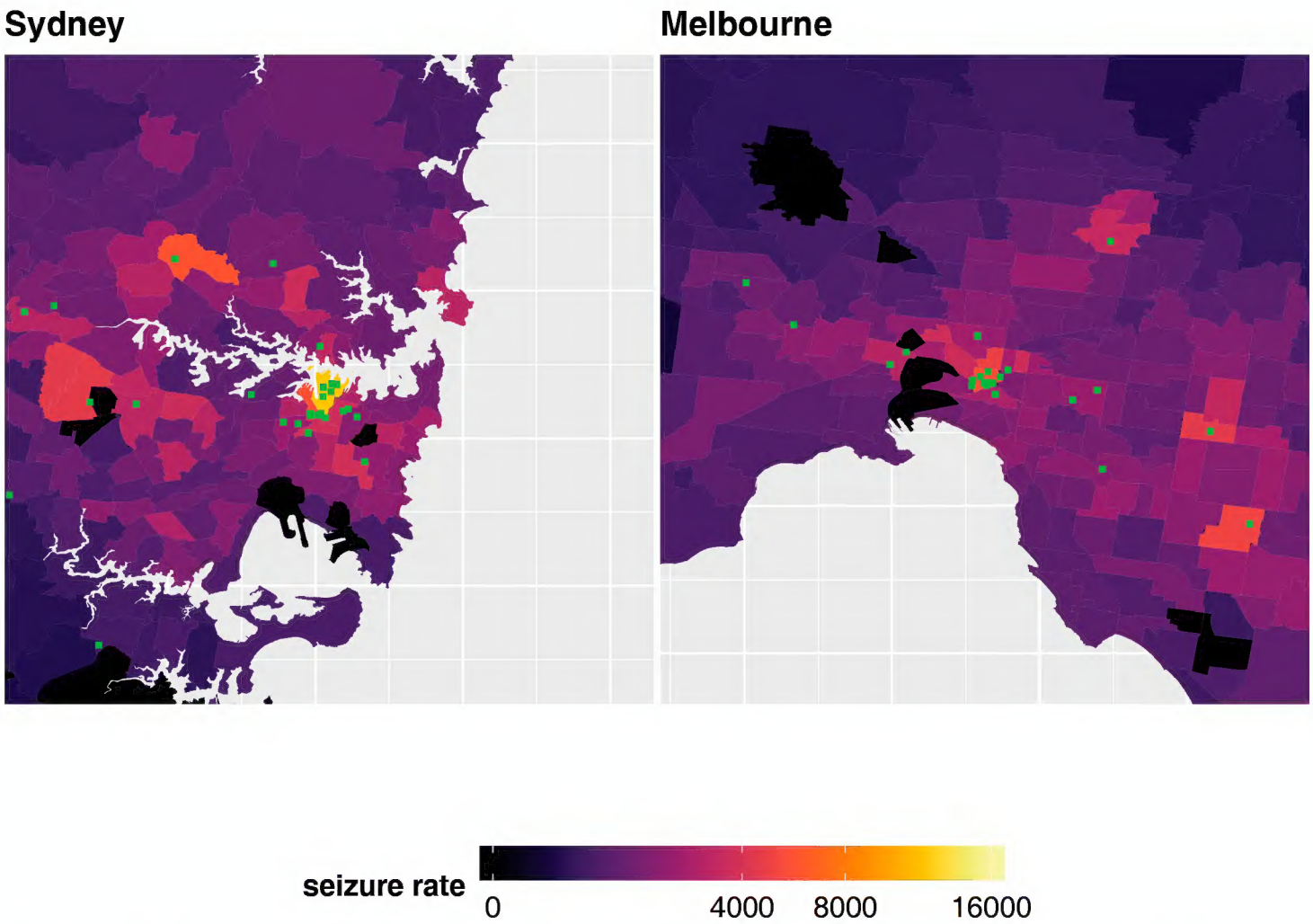

Figure 1. Seizure rate patterns by SA2 polygon for Sydney and Melbourne, per 100,000 people from 2008-2011. Green squares correspond to locations of university campuses.

seizures can be classed. Excluding live animals due to small counts, separate random forests were constructed for the seizures in each category, to compare the important demographic fields. The methods used were otherwise identical to those used for the analysis of the overall seizure rate, with the same relative variable importance measures available for interpretation.

Fields relating to no household income featured highly for all commodity categories, with university study and recent arrival also dominating most categories. East Asian Language A and/or some Country A ancestry were related to all the food categories (Fruit products, Legumes, Herbs, Mushrooms and Vegetable products), whereas another East Asian country was related to the destinations of animal products, an Eastern European ancestry and language were related to biologicals (this general class includes animal or microbial derived products such as foods, therapeutics, laboratory materials and vaccines) and language and ancestry from the Horn of Africa were related to human therapeutics. Seizures due to contaminated materials or soil/mineral samples were conspicuous for having no language or ancestry fields identified as important, which may be to do with the broad nature of these categories, with the biosecurity risk typically unrelated to the imported item itself and, instead, related to the presence of additional contaminants. Contamination, soil/mineral presence and plant products were all related to attendance of a university or other tertiary institution, with plant products also associated with those with high English language proficiency in particular.

Discussion

Our analysis provides insight into the spatial pattern of which, in the absence of border intervention, goods that present biosecurity risk may arrive in Australia. Each interception represents a measure of biosecurity risk averted. Our analysis suggests that there are population-level patterns in the demography of areas that receive a disproportionately high amount of biosecurity risk material by the international mail pathway. For example, SA2 regions with relatively high presence of university students who speak Language A at home are statistically closely related to high rate seizures of international mail articles. This insight suggests education campaigns might be used, for example, to provide information materials to relevant student groups at universities to better inform students about Australia's biosecurity imperatives or directed to particular suburbs of metropolitan areas which are more likely to receive risky goods. Future work, using independent data, could formally test the conjectured relationship between arrival counts and proximity to Universities.

The model described here can also be used to locate and profile particular demographic groups (and their usual residential locations) who are associated with a specific seized imported item. These education campaigns could therefore focus on the particular commodity types that this cohort are associated with, such as food-related importation information in Language A or human therapeutic information in the relevant African languages.

The usefulness of the results for providing information on policy and operational strategies is limited by the frequency with which the analysis can be repeated. While the biosecurity seizure data for international mail articles is continuously available for analysis, the Australian Census population data used in conjunction with mail seizure data is updated at 5-yearly intervals. If the length of time to complete tertiary education of the international student groups identified as high risk is less than 5 years, this means that educational campaigns such as distributing food related importation information would need to be directed towards new cohorts of students who meet the demographic profiles identified in this analysis, as well as current students. The demographic characteristics of the population may change over the 5-yearly interval, although such changes tend to be over a longer period. This change would limit the usefulness of the profiles for deciding where and to whom information about importing particular commodity types should be provided.

Well-directed information campaigns would have the potential to increase awareness and understanding of Australia's biosecurity importation rules and the reasons for them, amongst those who may inadvertently import biosecurity risk material. However, information on its own is unlikely to affect any deliberate, criminal importation activity. As well as providing a focus for educational campaigns, this analysis could also provide information about operational strategies that could target incoming international mail for inspection at the major points of entry. For example, the department is using a number of strategies for biosecurity compliance and inspection in the mail pathway, including working with

other agencies to improve screening and sampling techniques to intercept high-risk materials in international mail based on profiling information (Department of Agriculture 2016b). The analysis in this study has provided information about some of these strategies.

There is also an encouraging finding in relation to biosecurity risk: there was no evidence that these materials were destined for areas with greater agricultural land use generally, nor greater employment in such industries, as neither agricultural land use nor employment featured high in the list of variable importance.

These results also indicate there is considerable utility obtained from the use of the random forests model, which can be used to predict seizures of biosecurity material in international mail for a given situation. That is, given a set of demographic characteristics, the random forests model can be used to estimate the likely number of mail seizures for a specific imported product.

A shortcoming of the analysis is the lack of information available concerning the rate at which mail articles arrive at each SA2 polygon. Hence, the higher count of seized articles for specific polygons could be due in part to a higher volume of mail. However, this is not a particularly important caveat from the policy point of view, where the magnitude of biosecurity risk material is the primary apprehension. A further shortcoming is that we have no way of knowing from the mail interception records whether an inspection was random or targeted. Therefore there is no design-based guarantee of unbiased parameter estimates. The interpretation of the results should take this caveat into account.

Finally, it is in the nature of a geographical analysis, such as that outlined in this paper, that mail items that are posted to fake addresses could not be considered in the analysis. Unfortunately, the technique of using a fraudulent address is known anecdotally within the Department to be associated with a greater risk of deliberate, criminal importation. However, there already exist department procedures to monitor and analyse mail of this kind. This nature of risk is different from the kind of importation behaviour that can be targeted with the campaigns that would be motivated by this analysis.

Acknowledgements

The authors would like to thank the Australian Department of Agriculture and Water Resources for providing the data and motivation for this study. This research was supported by the Centre of Excellence for Biosecurity Risk Analysis, CEBRA.

References

Australian Bureau of Statistics (2016a) Australian Statistical Geography Standard: Volume 1 – main structure and greater capital city statistical areas. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.001

Australian Bureau of Statistics (2016b) Australian Statistical Geography Standard: Volume 5 – remoteness structure. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.005

Bivand R, Keitt T, Rowlingson B (2014) rgdal Bindings for the Geospatial Data Abstraction Library. R package version 0.8-16. http://CRAN.R-project.org/package=rgdal

Craik W, Palmer D, Sheldrake R (2017) Priorities for Australia's biosecurity system, an independent review of the capacity of the national biosecurity system and its underpinning intergovernmental agreement. http://www.agriculture.gov.au/biosecurity/partnerships/nbc/intergovernmental-agreement-on-biosecurity/igabreview/igab-final-report

Department of Agriculture (2016a) Biosecurity compliance plan 201617: Our plan for managing compliance. http://www.agriculture.gov.au/SiteCollectionDocuments/biosecurity-compliance-plan.pdf

Department of Agriculture (2016b) Biosecurity compliance statement. http://www.agriculture.gov.au/SiteCollectionDocuments/biosecurity-compliance-statement.pdf

Department of Agriculture (2018) Biosecurity Import CONditions database. https://bicon.agriculture.gov.au/BiconWeb4.0/

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15(1): 3133–3181.

Greenwood PE, Nikulin MS (1996) A guide to chi-squared testing, volume 280. John Wiley & Sons.

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer. https://doi.org/10.1007/978-0-387-84858-7

Kahle D, Wickham H (2013) ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3.

Kancans R, Stenekes N, Benedictos T (2010) Improving engagement of culturally and linguistically diverse persons in agriculture, fisheries and forestry. http://data.daff.gov.au/data/warehouse/pe_abarebrs99000003/ImpEngag.pdf

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22.

Mascaro J, Asner GP, Knapp DE, Kennedy-Bowdoin T, Martin RE, Anderson C, Higgins M, Chadwick KD (2014) A tale of two forests: Random forest machine learning aids tropical forest carbon mapping. PloS One 9(1): e85993. https://doi.org/10.1371/journal.pone.0085993

Meyerson LA, Reaser JK (2002) Biosecurity: Moving toward a comprehensive approach a comprehensive approach to biosecurity is necessary to minimize the risk of harm caused by non-native organisms to agriculture, the economy, the environment, and human health. BioScience 52(7): 593–600. https://doi.org/10.1641/0006-3568(2002)052[0593:BMTACA]2.0.CO;2

Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Computers & Geosciences 30: 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Pebesma EJ, Bivand RS (2005) Classes and methods for spatial data in R. R News 5(2): 9–13.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New York.

Maller C, Kancans R, Carr A (2007) Biosecurity and small landholders in peri-urban Australia. http://data.daff.gov.au/data/warehouse/brsShop/data/biosecurity_and_small_landholders.pdf

Supplementary material 1

Variogram of the residuals of the random forest model

Authors: Sandy Clarke, Nyree Stenekes, Robert Kancans, Chris Woodland, Andrew Robinson

Data type: statistical data

Explanation note: A variogram designed to visually assess whether there is spatial correlation present in the residuals of the random forest model. This does not indicate any spatial correlation.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/neobiota.40.28840.suppl1